# Homework: Map-Reduce

This document defines the **homework assignments** for the ["Algortihms" course @ Software University](https://softuni.bg/trainings/1194/Algorithms-September-2015). Please submit a single zip / rar / 7z archive holding the solutions (source code) of all below described problems.

## Install Apache Hadoop Working Environment

Install, configure and run **Apache Hadoop** instance on your laptop or in some cloud environment. You may use some of these approaches:

* **Local Hadoop installation** on Linux or Windows: follow the [installation guide](https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/SingleCluster.html).
  + Download and install **Java** SDK: <http://www.oracle.com/technetwork/java/javase/downloads/>.
  + Download and install **Hadoop**: <http://hadoop.apache.org/releases.html>
  + Configure system environment variables and paths: JAVA\_HOME, HADOOP\_BIN\_PATH and PATH.
  + In Windows, compile or download the missing files: bin\winutils.exe and bin\hadoop.dll.
* Download a pre-installed **Hadoop virtual machine** (VM)
  + [Hortonworks Sandbox](http://hortonworks.com/products/hortonworks-sandbox/)
  + [Cloudera QuickStart VM](http://www.cloudera.com/content/www/en-us/downloads/quickstart_vms/5-4.html)
* Provision a ready-to-use **Hadoop cluster in the cloud**:
  + [Amazon EMR](https://aws.amazon.com/elasticmapreduce/), [Azure HDInsight](https://azure.microsoft.com/en-us/services/hdinsight/), [Google Cloud](https://cloud.google.com/solutions/hadoop/), [Cloudera](http://www.cloudera.com/), [Rackspace](http://www.rackspace.com/cloud/big-data), [HP Cloud](http://www.hpcloud.com/solutions/hadoop)
  + Most cloud providers will request a credit card number, even for a trial subscription.

## Map-Reduce-Based CSV Report Analysis

You are given a potentially **huge CVS file** (terabytes) in the same format like the **real-estate sales report** from the samples (real-estates.csv). Write map-reduce functions (in Java or other language) to **calculate the total sum by city and type**. This means to extract each combination of **city + type** from the CSV file and calculate the **sum of prices** for it. The **output** might look similar to this:

* ANTELOPE / Residential 🡪 7 557 381
* ANTELOPE / Condo 🡪 115 000
* AUBURN / Condo 🡪 260 000
* AUBURN / Multi-Family 🡪 285 000
* AUBURN / Residential 🡪 504 000
* CAMERON PARK / Condo 🡪 119 000
* …

**Hints** to implement the map-reduce functions for processing the CSV file:

* The **mapper** function converts an **input CSV line** to a key-value pair.
  + The **input value** is a CSV text line which needs to be split and parsed to extract the columns holding the city, type and price.
  + The **output** **key** holds **city + " / " + type**.
  + The **output** **value** holds the **price**.
* The **reducer** function **accumulates the prices** for each key.
  + The **input key** holds the **city and type**.
  + The **input value** holds a **sequence of prices**.
  + The **output key** is the same like the input key.
  + The **output value** is the sum of all prices.